Gene 438 (2009) 17–25

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Conservation of selection on *matK* following an ancient loss of its flanking intron

Aaron M. Duffy [a,*], Scot A. Kelchner [b], Paul G. Wolf [a]

[a] *Department of Biology, Utah State University, Logan, UT 84322, USA*
[b] *Department of Biological Sciences, Idaho State University, Pocatello, ID 83209, USA*

ARTICLE INFO

Article history:
Received 10 October 2008
Received in revised form 29 January 2009
Accepted 10 February 2009
Available online 21 February 2009

Received by G. Theissen

Keywords:
Group II intron
Secondary structure
trnK
Intron-encoded protein
Fern

ABSTRACT

The chloroplast gene *trnK* and its associated group II intron appear to be absent in a large and ancient clade that includes nearly 90% of fern species. However, the maturase protein encoded within the intron (*matK*) is still present and located on the boundary of a large-scale inversion. We surveyed the chloroplast genome sequence of clade-member *Adiantum capillus-veneris* for evidence of a still present but fragmented *trnK* intron. Lack of signature structural domains and sequence motifs in the genome indicate loss of the *trnK* intron through degradation in an ancestor of the clade. In plants, *matK* preferentially catalyzes splicing of the *trnK* intron, but may also have a generalist function, splicing other group II introns in the chloroplast genome. We therefore tested whether a shift in selective constraint has occurred after loss of the *trnK* intron. Using previously unavailable sequences for several ferns, we compared *matK* sequences of the intron-less fern clade to sequences from seed plants and ferns with the intron and found no significant differences in selection among lineages using multiple methods. We conclude that *matK* in ferns has maintained its apparently ancient and generalized function in chloroplasts, even after the loss of its co-evolved group II intron. Finally, we also present primers that will allow amplification and nucleotide sequencing of the phylogenetically useful *matK* gene in additional fern taxa.

1. Introduction

The plant chloroplast gene *matK* has long sparked the interest of molecular evolutionary biologists. Its open reading frame is associated with a group II intron (Fig. 1A) that interrupts the coding sequence of tRNA^{Lys}(UUU) and it shows a much faster rate of sequence evolution than many other chloroplast genes (Wolfe et al., 1992; Hilu and Liang, 1997). The discovery of *matK*'s presence in the highly reduced plastomes of nonphotosynthetic plants (Wolfe et al., 1992; Ems et al., 1995) surprised many researchers and pointed to its probable role as a maturase that catalyzes the splicing reactions of more than one group II intron in the chloroplast genome (e.g., Ems et al., 1995; Liere and Link, 1995; Vogel et al., 1999).

Many known group II introns possess their own intron-encoded protein (IEP), which assists in splicing its host intron (Toor et al., 2001; Hausner et al., 2006). However, nearly all of the 20 or so group II introns in plant plastomes show severe degradation of their maturase open reading frames. This condition suggests that plant chloroplast group II introns no longer need to maintain their own splicing co-factor, an observation that many consider to be a strong indication of *matK*'s role as a generalist maturase (reviewed by Hausner et al., 2006).

The purported generalist function of *matK* would be unusual for an intron maturase. Introns and their IEPs are thought to have co-evolved: similar phylogenetic relationships are found among IEP sequences as are found among their intron RNA structures (Toor et al., 2001). Hence, there is a strong likelihood that the two components of a complete intron sequence (the intron itself, and its IEP open reading frame) are indelibly linked in terms of structure and function. Any shift in the intron's primary sequence or secondary and tertiary structure would likely correlate with a change of its IEP sequence (and function) if it is to successfully pass through the filters of natural selection.

Interestingly, the catalysis link between host intron and IEP continues to be strong in *matK*, even though it may also function as a generalist splicing co-factor for many chloroplast introns. Vogel et al. (1997, 1999) have shown *in vivo* that *matK* is required for *trnK* intron splicing in barley, and it will preferentially catalyze this reaction over the presumably less specific splicing of additional chloroplast introns, particularly those of structural subclass IIA (Liere and Link, 1995).

The unusual role of *matK* led us to question how selective constraints might vary for *matK* sequences in cases where the gene is no longer associated with its principal target, the *trnK* intron. This condition was observed in the chloroplast genome of the fern *Adiantum capillus-veneris,* which possesses *matK* and shows evidence of *matK* transcription (Wolf et al., 2004), yet appears to lack *trnK* and

Abbreviations: IEP, intron-encoded protein; ORF, open reading frame; CTAB, Cetyl trimethylammonium bromide; nt, nucleotide; RT, reverse-transcriptase; mya, million years ago.

[*] Corresponding author. Tel.: +1 435 797 0052; fax: +1 435 797 1575.
E-mail address: aduffy@biology.usu.edu (A.M. Duffy).

0378-1119/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2009.02.006

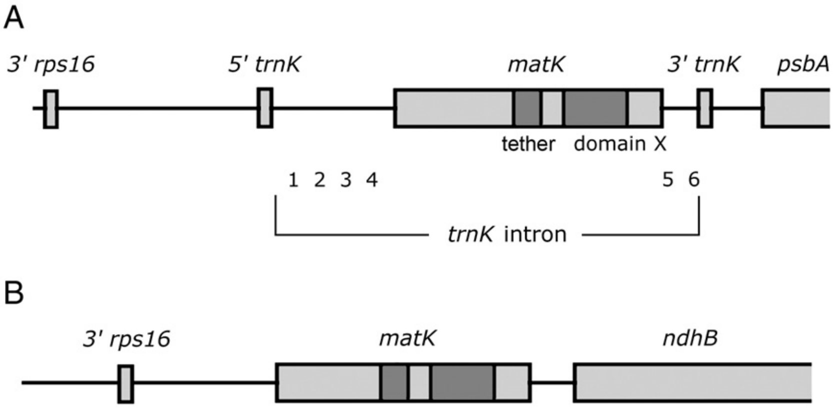


Fig. 1. The *trnK*/*matK* region as found in most angiosperms (A) and in the fern *Adiantum capillus-veneris* (B) roughly to scale. Transcription proceeds from left to right. The tether and domain X are relatively conserved functional regions of the *matK* protein. Numbers 1 through 6 designate the location of domains 1 through 6 of the *trnK* intron, respectively.

its intron (Wolf et al., 2003; Fig. 1B). Genome mapping studies indicate that the loss of *trnK* and its intron is associated with an ancient inversion event in the ancestor of a large clade of leptosporangiate ferns (Hasebe and Iwatsuki, 1992, Stein et al., 1992; Roper, 2007). This is an old lineage (~265 mya), which includes nearly 90% of the approximately 11000 extant fern species (Pryer et al., 2004). This genome structure would also explain the failure to obtain *matK* sequence from ferns: the PCR primers used in other plants are located in the missing *trnK* exons (e.g. Hilu and Liang, 1997; Hilu et al., 2003; Hausner et al., 2006).

Cases of *trnK* intron loss with *matK* retention have been inferred only twice before in plants: once in the highly reduced chloroplast genome of the achlorophyllous parasitic plant *Epifagus virginiana* (Wolfe et al., 1992; Ems et al., 1995) and once in the chloroplast genome of *Cuscuta reflexa*, a parasitic plant with reduced photosynthetic activity (Funk et al., 2007). However, these are recent changes that may not be shared with other extant taxa of each lineage, whereas the fern example is likely due to a very old event resulting in an intron-less *matK* in the majority of extant fern species. Ferns therefore present an opportunity to study the possible shift of selective constraints on an IEP (*matK*) after isolation from its co-evolved intron (the *trnK* intron) in a well-sampled and ancient clade of plants.

In this study, we use computational methods to establish whether the *trnK* intron is indeed absent from the *A. capillus-veneris* chloroplast genome or is instead divided but still functional, by searching for conserved intron sequence elements and intron-specific secondary structures. We then test whether any of *matK*'s major protein domains have experienced a shift in selective constraints after the loss of the *trnK* intron. We did this by: (1) obtaining *matK* sequences for additional fern taxa with and without a contiguous *trnK* intron, (2) comparing patterns of nucleotide and amino acid conservation across *matK* sequences in ferns and also seed plants, and (3) comparing rates of nonsynonymous to synonymous nucleotide substitutions in these groups using several methods. We also present primers for amplifying and sequencing a portion of the *matK* gene in ferns that are missing the *trnK* intron.

2. Materials and methods

2.1. Search for trnK intron in Adiantum

The *trnK* intron is not present in its expected location in the *A. capillus-veneris* chloroplast genome (Wolf et al., 2003), although this observation alone does not confirm its complete absence. Recombination involving group II introns has led to many known cases in which intron fragments are dislocated in a genome yet retain their function through *trans*-splicing mechanisms (Chapdelaine and Bonen, 1991; Bonen, 1993; Ems et al., 1995; Knoop et al., 1997; Jarrell et al., 1988; Malek and Knoop, 1998; Qiu and Palmer, 2004). Hence, it is important to our study that we first confirm an absolute loss of a conserved *trnK* intron sequence in the *A. capillus-veneris* chloroplast genome prior to interpreting any shift in patterns of selection among *matK* sequences that might be indicative of a change in the protein's function. We devised a method that first tests for the uniqueness of sequence within highly conserved structures of the *trnK* intron and then identifies several small sequence elements that would assist in locating intron structural fragments within the *A. capillus-veneris* chloroplast genome.

We constructed estimated secondary structure models for the *trnK* intron/*matK* sequence regions from *Pellia borealis*, *Marchantia polymorpha*, *Sphagnum platyphyllum*, *Cycas panzhihuaensis*, *Pinus thunbergii*, and *Atropa belladonna* (Supplementary Table 1; see Hausner et al., 2006 for additional *trnK* intron secondary structure models) using the domain-by-domain folding strategy of Kelchner (2002). The resulting RNA structures were then compared to identify conserved structural elements. Sequences of these structural elements (Table 1) were used to search the intergenic spacer regions immediately upstream and downstream of the *matK* ORF in *A. capillus-veneris* and all positive matches were explored using localized RNA folding by Mfold (Zuker, 2003) to survey for possible group II intron secondary structures. Sequences from both the upstream and downstream noncoding regions surrounding the *A. capillus-veneris matK* ORF were also folded with Mfold using an arbitrary “sliding window” approach with multiple sequence lengths shifting outward from the *matK* region in an attempt to identify helices that could be homologous with *trnK* intron domains. In particular, we looked for evidence of *trnK* intron domains 1 through 4, which would likely be proximal upstream of the *matK* ORF (Fig. 1).

The nucleotide sequence of domain 5 of the *trnK* intron is relatively well conserved among land plants (Supplementary Fig. 1) and should be recognizable if it is present in the chloroplast genome. The 34 nucleotides of domain 5 from *P. borealis*, *M. polymorpha*, *S. platyphyllum*, *C. panzhihuaensis*, *P. thunbergii*, and *A. belladonna* were used individually for BLAST searches against the GenBank database to verify

Table 1
Sequence elements used in searches for *trnK* intron fragments in *A. capillus-veneris*.

5′ nt	D2 5′	D2 3′	EBS2	EBS1	D3 3′	D4 5′	D4 3′	D5 5′	D5 loop
GTGCG	GGAT	ATCC	AGCTT	GTTAGAA GTTAAAA	ATCGC ATTGC	ATGTA ATGCA	ATCGC ATTGC	GCCG	GAAA GAGA

All matches identified in the *A. capillus-veneris* chloroplast genome were subjected to RNA folding to search for *trnK* intron fragments. Elements correspond to highly conserved *trnK* intron structures in a liverwort (*Pellia borealis*), a moss (*Sphagnum platyphyllum*), a conifer (*Pinus thunbergii*), and a flowering plant (*Atropa belladonna*). Nomenclature follows Michel et al. (1989).

that these nucleotides would identify domain 5 in other land plant *trnK* sequences, against the *Angiopteris evecta* chloroplast genome sequence to verify that they would identify domain 5 in a fern, and directly against the *A. capillus-veneris* chloroplast genome sequence as an attempt to locate *trnK* domain 5 in this genome. Each BLAST (blastn) search employed a low complexity filter with an expect threshold of 10 and word size of 11, and the return limit was set at the maximum 1000 matches from a Eukaryota virtual database.

2.2. DNA extraction, amplification, and sequencing

Taxa were chosen to represent most major fern lineages (Supplementary Table 1). Genomic DNA for new sequences was extracted using Qiagen DNEasy kits or a CTAB method (Doyle and Doyle, 1987). Published seed plant *matK* primers failed to amplify *matK* in fern taxa, and attempts to design universal fern *matK* primers using the *matK* sequences of *A. capillus-veneris* and *A. evecta* were not successful due to the lack of highly conserved regions. Sequences for *matK* in *Osmunda cinnamomea*, *Marsilea mutica*, and *Dicksonia antarctica* were obtained by amplifying large fragments using primers in flanking genes and primer walking. Using these sequences and the *matK* sequences of *A. capillus-veneris* and *A. evecta*, two sets of fern *matK* primers were designed: one set for the more recently derived "modern" ferns and one set for the more basal "early" ferns (Fig. 2). These primers were used to amplify and sequence portions of *matK* from *Lygodium japonicum* and *Pteridium aquilinum*. Because these primers are located within *matK*, to obtain the complete *matK* sequences, flanking regions were amplified and sequenced using these primers combined with primers in flanking genes. We also used the published *matK* sequences of the lycophyte *Huperzia lucidula*, the seed plants *Pinus koraiensis*, *Amborella trichopoda*, *Nymphaea alba*, *Magnolia dealbata*, *Helianthus annuus*, *Triticum aestivum*, and *Cycas pectinata*, and the monilophytes *Psilotum nudum*, and *Ophioglossum petiolatum* from GenBank (Supplementary Table 1). The *matK* sequences of *Isoetes engelmannii* and *Equisetum arvense* were extracted from their unpublished complete chloroplast genomes (K. Karol, personal communication).

2.3. Phylogenetic analysis

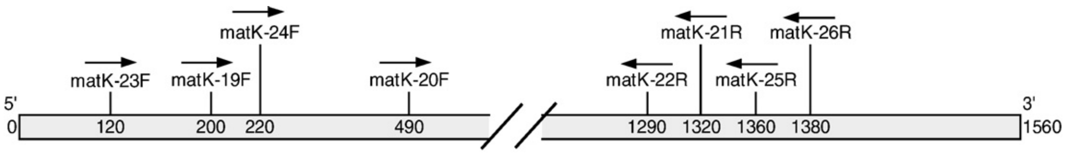
Nucleotide sequences were converted to amino acid sequences and aligned using clustalW with manual refinement informed by the nucleotide sequence and following the alignment principles of Kelchner (2000). The alignment was then converted back to nucleotide sequences for subsequent analyses. A single nucleotide gap was opened after nt 149 in *Huperzia* to shift the reading frame so the rest of the sequence would align. To match known RNA editing in *Adiantum* (Wolf et al., 2004), ACG (T) was changed to ATG (M) at nt 5, TCA (S) was changed to TTA (L) at nt 497, TCC (S) was changed to TTC (F) at nt 1001, and CCG (P) was changed to CTG (L) at nt 1004. RNA editing, including editing to repair stop codons, appears to be common in the monilophytes (Wolf et al., 2004) and *matK* in general (Barthet and Hilu, 2007). Several of our sequences contain stop codons that we changed to extend the open reading frame: TAA was changed to CAA (Q) at nt 559 in *Isoetes*, TGA was changed to CGA (R) and TAA was changed to CAA (Q) at nt 256 and 262 of *Dicksonia* and at nt 262 and 268 of *Pteridium*, and TGA was changed to CGA (R) at nt 250 and 256 of *Marsilea*. Making these changes extended the ORFs to approximately the same length as other *matK* sequences. Sections of the alignment containing indels that could not be aligned unambiguously were not included in subsequent analyses (Fig. 3).

We performed Bayesian Metropolis coupled Markov chain Monte Carlo analyses using MrBayes v3.1.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003). The data were partitioned by codon position and each partition was assigned its own model of nucleotide substitution (GTR + I + gamma) as determined using MrModeltest 2.2 (Nylander, 2004), a modified version of Modeltest 3.6 (Posada and Crandall, 1998). Using a random starting tree, we performed three separate runs with four chains each for 1 250 000 generations, sampling every 1000 generations. We plotted the log probability of observing the data by generation to detect stationarity and discarded the first 250 samples as "burnin." We pooled the post-burnin trees from each run and calculated a majority-rule consensus tree.

2.4. Tests of relaxed selective constraint

We compared the patterns of nucleotide and amino acid conservation along the length of the *matK* sequence in ferns with the *trnK* intron, ferns without the *trnK* intron, and seed plants. By focusing on the pattern of conservation rather than on absolute levels of conservation, we reduce bias due to different sample sizes, relationships between the groups, and whether taxa are equivalent between groups.

Several methods are available for detecting shifts in selection using the ratio of nonsynonymous to synonymous substitution rates (d*N*/d*S*; reviewed by Yang and Bielawski, 2000). If *matK* has been under relaxed constraint in the chloroplast genomes of ferns since the loss of *trnK* and its intron, we should be able to detect an increase in d*N*/d*S* ratios relative to ferns that have retained *trnK* and its intron. Those differences might be expected to be most evident in the active regions of the gene: domain X, which is associated with splicing activity, and the "tether region," which represents part of a reverse-transcriptase (RT) domain (Mohr et al., 1993; Hausner et al., 2006; Fig. 1). We tested for shifts in selection specifically on these domains, as well as on the gene as a whole. However, when using these much smaller subsets of the data, no differences in the pattern of substitution rates among the lineages were found and no statistical tests found significance that was not also present for the entire *matK* sequence, so only the results for the entire *matK* sequence will be presented.



"modern" fern primers:	
matK-19F	CGGCGAAACGTTTAATYGRTAG
matK-20F	CTTCATTCAGAGACTYYAATTCG
matK-21R	ACAGGCTAAMGTAYTATCACATG
matK-22R	ATATATTTCAATCTRCGCAAYC

"early" fern primers:	
matK-23F	AAGATGATTCTTACKCRATTGCT
matK-24F	TAATTGATAGGATRCGTCARCAAG
matK-25R	CGTGTTGTACTTTTRTGTTTACARGC
matK-26R	GTTTACAGGCYAARGTTTTATCGC

Fig. 2. "Universal" fern *matK* primers and their approximate annealing positions within a representative fern *matK* sequence. "Modern" fern primers were designed to anneal to regions conserved within *Marsilea*, *Dicksonia*, and *Adiantum matK* sequences. "Early" fern primers were designed to anneal to regions conserved between *Angiopteris*, *Osmunda* and *Marsilea*.

```
             1          11         21         31         41         51         61         71         81         91         101
             |          |          |          |          |          |          |          |          |          |          |
Huperzia     RIDKLQTKRF LYLLLFQEDI HAIACNRFSK KDSSLSFITR KRLISKIRQQ NLLEGGGHQS FDESIREVLT SVLQIVEEIL NSFQSIHSIF PSMEDHFSHR HHFSDIKIPY
Magnolia     MEELQEIQD. ..P..L..Y. Y.L.HDHGLN SNFGY.SLIV ....TRMH.. .H.ISVNDRF VGQMVS.GFA VIME.L..KS HNLR...... .FF..KL..L N.V...L..H
Adiantum     KSGP.GENC. .HP...L.NF YLTDGK.R.H ALVFG.TVAV ....GSV.DP .F.KSVSVDG L.HPLLKMTY LI.G.PA.TS KML.....M. LFL..R..KS N.ILEADL.H
Pteridium    KF.A.EED.. ..TF..L..F YS..RK.CLD PLVFG.AVAT ..S.GG..D. .YS.SEFVGR .NHGLL.TIC LI.E.DVKKS KFS....LL. LFL..RLPKS NQVL.TE..Q
Dicksonia    KLGT.KREC. ..PFF.KD.F YS...K.P.D PSVFG.M.SA ....DRM.H. DYSDSESGS. SNYALL.TIC L..E.SMKRW KPS....... LFL..RLPNS N.VL.T...Q
Marsilea     KDGGVKKDF. ..T.F.K..S YS..YK..PG SSVFG.I.SS ..I.DRM... DYSRSESGGR WNYALV.AIC L.SEKSTKKW KPS....... .FL..RLPNS T.LL.TR...
Lygodium     STSE.RI.C. .H.YV..D.. YPT.WTHL.G LNFHA.I.SV ......M... .YSKFEYDEN LGNALLKGIF IA.E.SL.KW K..R....ML LF...R.PQC NNIL.V..SH
Osmunda      .FE..RKEC. ..C..S.D.P Y...Y.C.LN SN.GF.I.SI ....KG.... D.SGFLPDR. GND.LLKGIA L..E.PMRGW K.L....... LF...R.L.S NYVL.T...Q
Angiopteris  K....KRQ.. ..TIP..... Y...YT.S.. SN.V..V..I ....TQL... .F.KNQLDHP YNKLLL.G.I LI.EVILGRW K.L.....L. LF..NK.L.S NFIL.....Q
Equisetum    LFENIV.Q.. ..P...HKEF YVVTSDS.ID SEFV...LVI R...NR..NL .NSKFIQTK. ICFYLLKA.N M..ETIKRVW K.YK..F.TC SFV.EK.IFS NQIL.L...H
Ophioglossum V.GG.RSRY. S.P...HD.P ..F.R..S.D SYRNR.ILAV .....RM.RR I..RLRN.PY T.RYLH.GVA V..EFSRSLW .CIRTL...P LF..HRYLNS NLVL.LR...
Psilotum     K.G.VKSLWQ ..P...R..L Y...Y..S.S LHRDS.LVII ....NR..N. TTSKFQ.SLK INR.LF.G.. VLFEM.QYQW ..L....... LF...R.FYS NSILGL....

             111        121        131        141        151        161        171        181        191        201        211
             |          |          |          |          |          |          |          |          |          |          |
Huperzia     FIHPEILIRI FRRRIQDAPF SHLLRFVFYE YQDKSWKEIS KLSTFLWNYY VHEFESTLVS LWKRTFHFNS FPLLDRTQSI RKVKHIEKLL SGKKYSIHYV RYENNLMIAL
Magnolia     P..L...VQT LHCW.....S L....LFLH. .RK.FS..NQ RFFLL.Y.SH .Y.C..V..F .R.QSS.LR. TSF.E..HFY G.IE.LVVVK TL.DPFM... ..QGKSIL.S
Adiantum     NL.L.T...L ...Q.K.VS. L....I..RK RKKTPGGGQG SVDIPVR.F. IF.ID.L.LI P..QVYK.RV NYIDSCNIIR KEIYASAYKK ASRSLW...G .WR.KFL..S
Pteridium    NL.L.TS..L ...Q.K.VS. .....I..HK .KKICKQKER SIDIL.R.F. IY.ID.L.LV ..R.MHESQP RYPDQNNIIR KEAHGSIYGA ADRSLC...G ..R.KSF..F
Dicksonia    NL.S.TP..M .......... P..S.LALHK .GKNWKGKKK NITIL...F. IY.I..LPLF ...QVYRSQ. RYTDQNNITR KERHAY.YGT ANRSLC.P.G ..K.HSL..F
Marsilea     NL.S.T...M .......... L.F..LLLSK .ITDPKL.VK ..ANM..... IY.I..S.LF .R.QLRGSQ. I.TGLSNTIR KERHSDWYGA INRSLY...A ..K.RSLV.V
Lygodium     LL.S.T...T ...QVR.VS. L.....R.I. SKKE.SI.EN I.FYL.L..H IS.M..V.L. ....LIPTSQ LK.RSDQN.. FQKDRGKLKP LNEGSCV..G ..K.QFLLVF
Osmunda      .T...AS..M ........S. L.P..L..H. ..KI.SR.KN S.AIL...F. AY.S..L.IP ....FSRLR. DSIF.QINLL ...E.VTESI LR.NPC...G ....HFIL..
Angiopteris  SL........ .......T.. L..S.SILH. ...TFSR.QN S.LIL..... .Y...YLV.. S...FSRLQ. IFRI...HFD ..I..VIRPI .S.NPC.... ..K.HSVL.F
Equisetum    .....SF... L.QQ.K..S. L..T..FVH. .KE.FN.RNK FVTFSWNFF. IF.L.FF.T. .LT.FINNLV LS.F.QINLL E.INNNN.SK RIQNSC.... ..Q.HCIM.S
Ophioglossum LL.T.....L L....R.V.. L...GTISHN L.NTPCVKNN R.FLV..... SY...NL..L IG..FSQLR. I..V..IHYH .R..RVIR.. PSRNPC..F. ..G..CIL.F
Psilotum     YF...MI..L .....K.VFL L..V.LL.HN ..ET.L.DSK R..IL.R.H. FY...NQ..P ....FVQLQ. LS.M.Q.NPL Y.M..GLGSP FLETPC.... ....HSI..F

             221        231        241        251        261        271        281        291        301        311        321
             |          |          |          |          |          |          |          |          |          |          |
Huperzia     KGTKFLVHKW KCYIIRFWQY YFHCWFKPCR VSPRESSKEC LTFLGYILGI RPQIIVVQAK MINNLPITSI ISKELCAIIP VFHSIKLLAR EKFCNTLGHP IGKLAWTTSE
Magnolia     ...HL.MK.. .SHLVH...C H.YL.SL.D. IHINQLYNHS .Y....LSSV .LNTS..RIQ .LE.SFLIDT SI.KFETLV. IIPL.GSV.K A....VS... .S.SVRAD.S
Adiantum     E..HYF.K.M LY.LWILLK. H..YRI.SNE PWIKLLPTS. VS....T.LA QLVSKN.RIE TVTD.Y.SIL GG.KFYPK.. NSII.TT..K QR..DFT.R. ...S..V..T
Pteridium    E..V.FAK.. MY.FLILLRS H..YRTEFDK MHLKLL.NG. IS..S.T.TV QSVSKN..IE TTTGFYMSFS CD.KFYPKA. ISLLV....K G...DST... VS....AVLA
Dicksonia    G..RYFAR.. I...LILVES HC.YRTESNQ MCIKLL.NG. VSL...TS.V QSG.KK.RVG ATEGSH.SLS .A...FPKV. ISLLV..M.K DN..DNT.R. VS.......T
Marsilea     S..RYFAR.. I.F.LLLIES Q..LSNESNQ .LVKI...N. ILL...T.DV QSIATKARVG T.GESYLSLF FA.KSF.KL. ISSIV..M.K .N..DSA... VS.SC...LP
Lygodium     T.IED.TD.. IFH.LMVI.. ...SRIHLR. INSKKL.TSS SLV.A.S... QSMSKK..VL FVGESYG... ...GFLPKV. TLPL.RFMEK .G.RDST... VSRSD..LLT
Osmunda      R..HSSAR.. IH.LLMLR.S HY...IQ.Y. ICIGRLPRN. FS....T..V .LI.KE.RVG TVDESY..AS .A..FRSK.. TLLL..S... .G..DSS.R. VSRS.R.ALT
Angiopteris  Q..NY.AK.. RN.LLN.... H....VQ.H. IFLKRF.RNS FS........ .TR.NK.... .EDE....CL .T....P... FLLLVNS... GG..TN..R. VS..S...LT
Equisetum    E.FY.HDTN. IY..LNI..F FM.L.IQ.F. F.TKHFQ.QS FF....QF.R ESKLLK.RSI SLDKS.TIYS RL.KNLLKTQ IVYP.DF..K .G..DIS.Y. .SRST....T
Ophioglossum E.SNYSAR.. IH.LL....C NH.S.LRTQ. IRIIK.YINS IF....T..S LSEMVGIK.. TMD..ST.H. TFRV..PKV. TSL..RS... .G...G..F. .SRS..A..T
Psilotum     ....SI.N.. IK.LVG.... NY.Y.LQ..Q IDI.RP.RR. FS.M.....F .SRM.K.HT. R.DESST.HC .I..F..S.. TSSL.ES.T. .G..DSS.R. V.RST..ILK

             331        341        351        361        371        381        391        401        411
             |          |          |          |          |          |          |          |          |
Huperzia     DDDIPNQFNH IWKNIFYYYS GCLNRNGLYQ IQYILRFSCA KTLACKHKST IRVVWKKYGS RLFPRYSYKE RHKKRFWYLD VIQ
Magnolia     .S..I.R.GR .YR.LSH.H. .SSKKQT..R .K....L... R...R..... V.AFL.RL.. EFLEELEEEQ V.RE.I.... I.R
Adiantum     ..K.IDGYVQ L.QVFSL..G ASM.QYR.RR LIFL.QM..D S...G..R.. ..LLRC.SNV A.NQISKFEL SSSR.V.RSS S.R
Pteridium    ....L.R.VQ ...ILYL.H. ASI..D..RR LR....L..D S...G..R.. ..FLQRRFDL LPKTFSKSGS DKNQ.V.H.S L.R
Dicksonia    ....L.R.VQ T.RIFSL.H. .SI..D..RR LK.......D N......... T.SLRRRFD. I.LDNKE.EM SNNQ.V.H.S .TR
Marsilea     ....LKR.I. V.RTLSL... .SI..D..RR SKH......D .......... T.SLRRRFD. VVSDNKL.GK SSNQ.V...N ITR
Lygodium     .S..SKR.V. L..IFSR... .SRS.D..RK LR.......D .......... ..LIRQ.FD. TFLKIKRGPH FSNQ..RD.. I..
Osmunda      ....L.R.LR ..RS.S.... .SID.D...R LR.......D .......... T.SIRNRF.. .IL..LEES. SNDR.I.C.. I..
Angiopteris  ....LKK.DQ ..RSVY.... .SI.NH..FR LR..F..... .......... T.I...RFSL SFLRSFKKP. L..R...... I..
Equisetum    .EE.LLN..K ...SFYF..G .LIKKDI..R .K........ ....R..... T......IV. D.SLSSRR.K ND.....SS. ITK
Ophioglossum .T.TT.R..R L...L.I... .SSGLG...R .R........ .......... ..A...RF.. .FNL.SVNPY P.N....C.. .S.
Psilotum     ....L.KYHQ ..GDLSC... .SFS.D..WR AK...QL... ....Q..... T...RNHF.L KFITTSKNPF F.R.N..C.. I.R
```

Fig. 3. Representative *matK* nucleotide alignment including all fern sequences, one seed plant, and one lycophyte sequence. Nucleotide sequences were translated to amino acid sequences and aligned using ClustalW. Regions with indels could not be aligned unambiguously and were removed from the alignment and are not shown. Shaded portions of the alignment represent the tether region including RT subdomains V, VI, and VII (amino acids 207–270) and domain X (amino acids 293–393).

We used the program package PAML (Yang, 1997) to evaluate models that allow ratios of nonsynonymous to synonymous substitution rates to vary among lineages in order to test whether levels of selective constraint vary among the lineages of ferns without the *trnK* intron, ferns with the *trnK* intron, and the rest of the tree. This method (Yang, 1998) requires *a priori* hypotheses of which branches on the tree vary, and models are compared using likelihood ratio tests. We constructed several models (Table 2) ranging from the simplest, with all 3 lineages assigned a single d*N*/d*S* ratio (Model A), to the most general, with each of the three lineages assigned a separate ratio (Model D). PAML and another program package, HYPHY (Kosakovsky Pond et al., 2005), were both used to evaluate a free branch model, which assigns a separate ratio to each branch of the tree.

We also performed HYPHY GA Branch Analysis, using a genetic algorithm to assign d*N*/d*S* ratios. Each branch on the phylogenetic tree was assigned to a d*N*/d*S* ratio class with the optimal number of classes determined from the data (Kosakovsky Pond and Frost, 2005). We then looked for patterns in the way branches were assigned to ratio classes that might suggest differences between lineages.

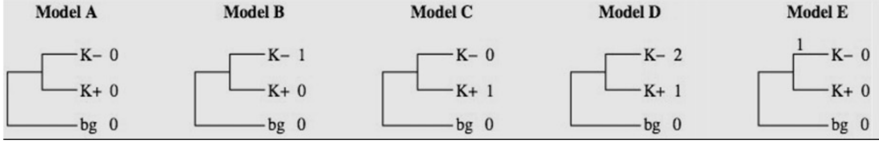
3. Results

3.1. Search for trnK intron in Adiantum

Using the domain-by-domain folding method on the *trnK* intron/*matK* sequence regions of land plants, we recovered core secondary structural models consistent with well supported group II intron models (Michel et al., 1989) and identified several conserved sequence elements related to structures shared across taxa (Table 1). Most of these short (4–7 nucleotides) elements returned positive matches when used in localized sequence searches upstream and downstream of the *matK* ORF in *A. capillus-veneris*, but none was

Table 2
Likelihood ratio tests of variation in d*N*/d*S* ratios between the ferns with the *trnK* intron (K+), the clade of ferns without the *trnK* intron (K−), and the rest of the tree (bg).

H_0	Assumptions	Models	chi^2	*df*	*P* value	What are we asking?
K− = bg	K+ = bg	A and B	0.03	1	0.8543	Does the K-minus clade differ from everything else?
K− = bg	K+ = free	C and D	3.05	1	0.0807	Does the K-minus clade differ from everything else if the K-plus ferns are free to vary?
root = bg	K+ = bg	A and E	0.22	1	0.6423	Does the single branch leading to the K-minus clade differ from everything else?
		A and Free	60.31	33	0.0026*	Does a model with separate values for each branch fit the data better than a one ratio model?



*$P<0.01$.
For each model, branches associated with lineages sharing a number (0, 1, 2) were constrained to share the same d*N*/d*S* ratio, while branches in lineages with different numbers were free to differ. For Model E, all branches were constrained to share a d*N*/d*S* ratio except the single branch rooting the ferns without the *trnK* intron (root).

embedded in sequences that could be folded into RNA secondary structures resembling those of the *trnK* intron from which the element was derived. We found no evidence of potentially homologous helical structures and their associated conserved sequence elements that would indicate a *trnK* intron fragment. We conclude that *matK* is no longer associated with a divided, but functional, *trnK* intron in the chloroplast genome of *A. capillus-veneris*.

BLAST searches using domain 5 sequences from each of six land plant taxa returned *trnK* intron sequences (but not other group II intron sequences) from nearly all the major plant groups including the fern, *Angiopteris evecta*. This is consistent with our expectation that the nucleotide sequence of domain 5 in the *trnK* intron is both unique to the *trnK* intron and highly conserved in land plant chloroplast genomes. These BLAST searches did not return a domain 5 sequence for the *trnK* intron in *A. capillus-veneris*. When these six domain 5 sequences were used for BLAST searches directly against the *A. capillus-veneris* chloroplast genome, only a few 11-nucleotide complements were returned with low expect values (0.045 to 1.9), but none demonstrated a secondary structure that was potentially homologous to domain 5 of a group II intron. None of the searches identified a sequence in *A. capillus-veneris* that resembled the *trnK* intron in either its primary or secondary structure, strongly favoring the hypothesis that the *trnK* intron, or recognizable fragments thereof, is no longer present in this chloroplast genome.

3.2. Phylogenetic analysis

The *matK* consensus tree topology (Supplementary Fig. 2) is similar to published phylogenetic hypotheses with just a few differences, primarily on branches that are not strongly supported in analyses using other genes. *Equisetum* is placed as sister to the rest of the monilophytes, rather than sister to *Angiopteris* as hypothesized by Pryer et al. (2004) and Qiu et al. (2006) or sister to the core leptosporangiates (*Marsilea*, *Dicksonia*, *Adiantum*, and *Pteridium*) as hypothesized by Magallón and Sanderson (2005). Within the seed plants, relationships differ somewhat from published hypotheses (Hilu et al., 2003; Magallón and Sanderson, 2005; Qiu et al., 2006). Despite these differences, the relationships for the comparisons we want to make are consistent with published hypotheses — the ferns and the seed plants form monophyletic sister groups, and the ferns without the *trnK* intron (hereafter referred to as the “K-minus clade”) form a well supported monophyletic group nested within ferns with the *trnK* intron (hereafter referred to as the “K-plus ferns”). The consensus tree topology could be biased by the small number of taxa included and also by the limitations of any single-gene phylogenetic analysis, so a constraint tree based on published topologies was used for subsequent analyses. Results of analyses using both topologies were compared to ensure that topology differences did not affect our inferences (data not shown).

3.3. Tests of relaxed selective constraint

The proportion of variable nucleotide alignment sites for K-plus ferns, the K-minus clade, and seed plants was lower than the percent total amino acid variation (K-plus ferns: 70% vs. 81%; K-minus clade: 67% vs. 82%; seed plants: 58% vs. 74%). This is consistent with values reported for other plant groups (Hilu and Liang, 1997) and suggests a lack of strong functional constraint on the gene as a whole since nucleotide variation typically translates into lower amino acid variation in functionally constrained genes (Graur and Li, 2000). However, any direct comparison of the percentage variation between K-plus ferns, the K-minus clade, and seed plants is probably inappropriate as these values are likely biased by the different sample sizes used and the phylogenetic relationships among the groups.

Whereas the percentage variation and absolute number of variable nucleotides and amino acids may be biased, the pattern of variation or distribution of relatively conserved or variable regions along the sequence can be compared directly. The pattern of nucleotide and amino acid variation along the length of the *matK* sequence is similar in K-plus ferns, the K-minus clade, and seed plants (Fig. 4), and is similar to the pattern described by Hilu et al. (2003) in the *matK* sequence for a large sample of angiosperms. The three groups have highly variable sequences but a region with less variability in one group tends to have less variability in the other groups as well. All three groups have regions of relatively conserved sequence located approximately 300 nucleotides (100 amino acids) from the 5′ end of our alignment, and approximately 200 nucleotides (65 amino acids) from the 3′ end of our alignment, with the latter region corresponding to part of domain X. These same low variability regions have been identified in other plant groups as well, including the nonphotosynthetic, *trnK*-lacking *Epifagus* (Hilu and Liang, 1997; Young and dePamphilis, 2000). All three groups also show a slight reduction in sequence variability within the tether region, but it is less conserved than the other two regions.

Likelihood ratio tests using PAML did not find significant differences between models varying in the levels of heterogeneity in the d*N*/d*S* ratio among lineages (Table 2, Table 3). The model with one ratio for the K-minus clade and one for the rest of the tree (Model B) does not fit the data significantly better than a single ratio model (Model A), suggesting that the d*N*/d*S* ratio for the K-minus clade is not different from the rest of the tree. This is the case even if the d*N*/d*S* ratio for the K-plus ferns is allowed to vary (Model C vs. Model D). To allow for the possibility that purifying selection on *matK* was

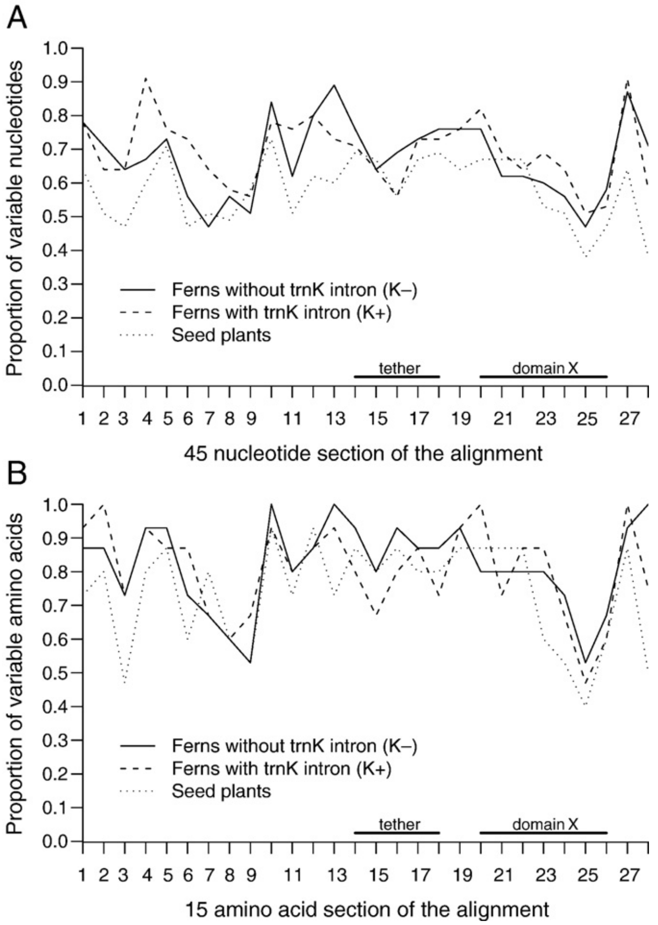


Fig. 4. Comparison of patterns of distribution of variable alignment positions for ferns without the *trnK* intron (K−), ferns with the *trnK* intron (K+), and seed plants. (A) Proportion of variable positions in each 45 nucleotide section of the *matK* nucleotide alignment. Each number on the *x*-axis represents a 45 nucleotide section of the alignment. (B) Proportion of variable positions in each 15 amino acid section of the *matK* amino acid alignment. Each number on the *x*-axis represents a 15 amino acid section of the alignment. Regions containing indels are excluded from the alignment. The approximate position of the tether region and domain X are labeled on the *x*-axis.

diminished following the inversion and loss of the intron but then increased again, we tested specifically for a change in d*N*/d*S* on the branch where the inversion that separated *matK* from *trnK* and its intron occurred (Model E). We tested whether a model with one ratio for that single branch and another ratio for the rest of the tree fits the data better than the one ratio model (Model E vs. Model A), but found no significant difference.

Table 3
Estimates of d*N*/d*S* ratios and log likelihood values for the ferns with the *trnK* intron (K+), ferns without the *trnK* intron (K−), and the rest of the tree (bg), under different models used in branch model tests.

Model	d*N*/d*S*	ln(Likelihood)	Parameters
A	bg = K+ = K− = 0.2899	− 18539.203	36
B	bg = K+ = 0.2913 K− = 0.2866	− 18539.186	37
C	bg = K− = 0.3188 K+ = 0.2187	− 18531.462	37
D	bg = 0.3399 K+ = 0.2185 K− = 0.2878	− 18529.936	38

E	bg = 0.2888
	Root = 0.3457
		− 18539.095	37

For Model E the single branch rooting the clade of ferns without the *trnK* intron (root) is compared to the rest of the tree.

We used a free ratio model to infer a separate d*N*/d*S* ratio for each branch of the tree (Fig. 5). The free model analysis was performed using both PAML and HYPHY, and though the two software packages gave slightly different values for d*N* and d*S*, the ratios were identical. Likelihood ratio tests show that the free ratio model fits the data significantly better than the one ratio model (Model A) (Table 2, Table 3). The d*N*/d*S* ratios for nearly all branches were less than one, suggesting mild purifying selection on the gene as a whole, consistent with tests of selection on *matK* in other taxa using protein side-chain composition (Barthet and Hilu, 2008). There is no obvious increase in d*N*/d*S* ratios on the branches associated with the K-minus clade. In general, values for the branches associated with the K-minus clade fall between those for the K-plus fern lineages and those for the seed plant lineages. There are three branches in the tree with d*N*/d*S* values greater than one, but these all have very short branch lengths so it is not clear whether the high ratios are valid or are due to chance. However, these large or infinite d*N*/d*S* values are not assigned to branches associated with the K-minus clade. These results do not suggest that the *matK* sequence of the K-minus clade is under relaxed selective constraint or positive selection relative to the other groups. If anything, the K-minus clade is under slightly stronger purifying selection than the seed plants and slightly weaker purifying selection than the K-minus ferns.

The HYPHY Genetic Algorithm Branch Analysis assigned each branch to one of four classes ranging from d*N*/d*S* = 0.139 to 0.531 (Fig. 6). None of the branches associated with the K-minus clade was assigned to the highest d*N*/d*S* class. All but one of these branches were assigned to the 2nd highest (and most common) class. Within all the ferns, the only two branches assigned to the highest d*N*/d*S* class were the branch rooting *Angiopteris*, *Equisetum*, and the Leptosporangiate ferns, and the branch leading to *Ophioglossum*. The rest of the

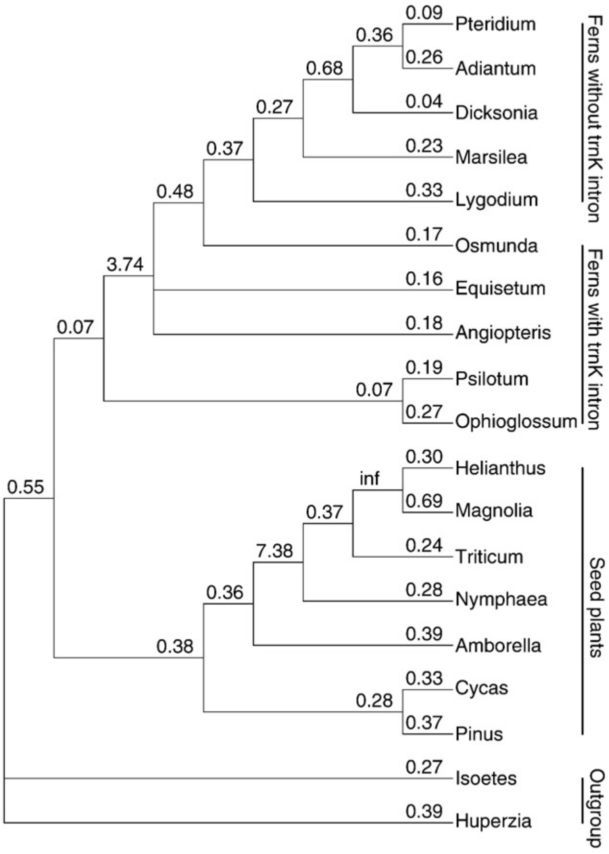


Fig. 5. Constraint tree with Free Model d*N*/d*S* ratio values with a separate value assigned to each branch. Tree topology is constrained to be consistent with published seed plant and monilophyte phylogenies.

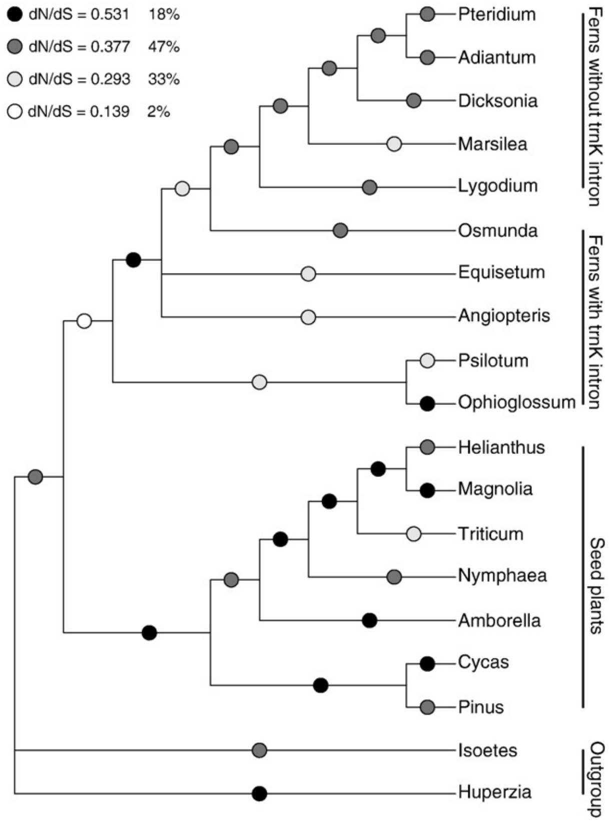


Fig. 6. HYPHY Genetic Algorithm Branch Analysis tree for the full *matK* sequence. Branches are assigned to d*N*/d*S* ratio categories with the optimal number of categories determined from the data. The percentages for the branch classes represent the proportion of the tree length (expected substitutions per site per unit time) evolving under that value of d*N*/d*S*. Tree topology is constrained to be consistent with published seed plant and monilophyte phylogenies.

branches allocated to the highest class were in the seed plants and outgroup. The only branch included in the lowest d*N*/d*S* class is the branch rooting the ferns. Overall, the class assignments mirror the results of the likelihood ratio tests, with the fern branches generally in the same or lower d*N*/d*S* classes than the seed plants and outgroup. Within the ferns, the branches for the K-minus clade tend to belong to higher classes than the K-plus fern branches, but still lower than those for the rest of the tree. None of the results suggests that there is relaxed purifying selection in the lineages where *matK* has been separated from *trnK* and its intron.

4. Discussion

Our investigation of the atypical *trnK*/*matK* condition in the *A. capillus-veneris* chloroplast genome (which is shared by other ferns in the K-minus clade) leads us to conclude that the *trnK* intron has been lost even though its IEP, *matK*, is retained. The position of the *matK* ORF at the border of an inferred inversion event had allowed for the possibility that the *trnK* intron was still functional in the genome as a *trans*-spliced, divided intron. If so, *matK* would be flanked by one or more of the *trnK* intron domains, and the highly conserved, diagnostic domain 5 sequence would be detectable in the genome. However, we found no evidence for any conserved *trnK* intron secondary structure or sequence motifs flanking the *matK* ORF, or for conserved *trnK* domain 5 sequence elements in the complete *A. capillus-veneris* chloroplast genome. These observations favor the hypothesis that the *trnK* intron was disrupted by an inversion event in an ancestor of the higher ferns, failed to persist by *trans*-splicing mechanisms in the chloroplast genome, and was subsequently degraded and lost.

If *matK* was functioning primarily as a splice factor for *trnK* and its intron, once *trnK* was lost, we would expect *matK* to be released from this particular selective constraint and to either accumulate substitutions and become degraded, or optimize to its role as a generalist catalyst of splicing IIA introns, or less likely, to gain new function. However, the similar distributions of conserved nucleotides and amino acids in the *matK* sequences of seed plants, K-plus ferns, and the K-minus clade suggest that substitutions are not being accumulated differently in the *matK*s that exist without the *trnK* intron. Additionally, analyses of d*N*/d*S* ratios using several methods – including the entire coding sequence as well as specifically focusing on the functional domains – do not support the hypothesis that *matK* is under relaxed selective constraint in taxa without *trnK* and its intron, as would be expected if the gene was no longer functional in these taxa. If anything, the d*N*/d*S* analyses suggest that *matK* in the ferns without the *trnK* intron has been under a similar level of selective constraint to that of the seed plants, whereas in the ferns with the *trnK* intron, *matK* is slightly more constrained.

The presence of *matK* cDNAs in *A. capillus-veneris* (Wolf et al., 2004) and another fern in the K-minus clade (Barthet and Hilu, 2007) suggests that *matK* continues to function despite the absence of the *trnK* intron. If *matK* were performing a new role in the genomes of ferns without the *trnK* intron, we would expect to see evidence of positive selection as the sequence responded to different selective forces, but there is no evidence of positive selection in any of the analyses. So, even after the loss of *trnK* and its intron, *matK* is apparently performing a largely similar function, consistent with the hypothesis that *matK* is involved in splicing other group IIA introns in the chloroplast (Vogel et al., 1999). Other genes in the *A. capillus-veneris* chloroplast genome with group II introns are *atpF*, *clpP*, *rpl2*, *rps12*, *trnA*, *trnI*, and *trnV* (Wolf et al., 2003; Funk et al., 2007).

The hypothesized inversion (Hasebe and Iwatsuki, 1992; Stein et al., 1992; Roper, 2007) that appears to have disrupted the intron is shared by members of a clade that diverged approximately 265 mya (Pryer et al., 2004) and includes nearly 90% of all fern species. This indicates that the *matK* condition in these taxa is both ancient and stable and contrasts with the other examples of inferred maturase isolation — the reduced chloroplast genomes of *E. virginiana* (Wolfe et al., 1992; Ems et al., 1995) and *C. reflexa* (Funk et al., 2007), which are members of much younger angiosperm clades.

Because the genomes of nearly 90% of all extant fern species do not have the typical *trnK*/*matK* arrangement, and many of the primers used to amplify *matK* in other plant groups are within *trnK* or its intron, ferns have been underrepresented in previous analyses of *matK*. The high rate of nucleotide substitution in *matK* makes it difficult to design primers that are conserved across large taxonomic groups, and taxon-specific primers are often required (Hilu et al., 2003). A recent proposal of two options for three-gene land plant DNA barcodes included *matK* in both options, but noted that additional work was needed to improve the performance of the primer sets (Chase et al., 2007). We have so far been unable to design a single set of primers to amplify *matK* across all the fern taxa used in this study, but designed two sets: one in sequence conserved across the early diverging fern lineages and one in sequence conserved across the more recently derived lineages within the K-minus clade (Fig. 2). There is some overlap in the taxa that each set will amplify, but so far the "early" fern primers have been tested successfully on *Osmunda*, *Lygodium*, *Gleichenia*, *Marsilea*, and *Dicksonia* and the "modern" fern primers have been tested successfully on *Adiantum*, *Dicksonia*, and *Pteridium*. Because these primers are located within the *matK* sequence they cannot amplify the entire *matK* gene. But, by combining these primers with knowledge of the gene order surrounding *matK* in ferns with and without *trnK* (Fig. 1), the entire *matK* sequence may be amplified using published *chlB* or *rps16* primers for the upstream flank and *ndhB* (K-minus clade) or *psbA* (K-plus ferns) for the downstream flank (Roper, 2007). The primers designed for this study should allow *matK* to be amplified and sequenced for most fern taxa.

4.1. Evolutionary significance of maturase isolation

Close functional association of introns and their encoded maturase is common among prokaryotic, protist, and fungal group II introns (Lambowitz and Perlman, 1990; Moran et al., 1995; Saldanha et al., 1999; Huang et al., 2005), many of which appear to have co-evolved with their IEPs (Costa et al., 1997; Toor et al., 2001; Zimmerly et al., 2001). Liere and Link (1995) established that *matK* binds preferentially to *trnK* intron pre-mRNAs in the mustard *Sinapis alba* when given the opportunity. Although *matK* is typical in that it maintains a tight functional association with its encoding intron, it is also unusual in its apparent contribution to splicing reactions of other group IIA introns (Vogel et al., 1999). One plausible history for the advent of *matK*'s generalist nature is given here. It is thought that mobile group II introns invaded ancestral chloroplast genomes sometime prior to the divergence of the charophytes and embryophytes (Toor et al., 2001; Sanders et al., 2003). As most extant mobile group II introns use IEPs for retrotranspositioning (Zimmerly et al., 2001), it is likely that each invading intron carried a maturase ORF in its domain 4 and one of these was the ancestor of *matK*. Consistent with this hypothesis are the observations that *matK* is always associated with the *trnK* intron when both are present, and that chloroplast genomes that lack group II introns also lack *matK* (Turmel et al., 1999; Lemieux et al., 2000).

After their establishment in plastid genomes, most chloroplast group II introns seem to have shifted from a specific interaction with their IEPs to a more relaxed association with a variety of splice-assisting proteins such as those encoded by nuclear genes *crs1* and *crs2* (Jenkins et al., 1997), plastid ribosomes (Hess et al., 1994), and other as yet to be determined factors (Vogel et al., 1999; Matsuura et al., 2001; Toor et al., 2001). Such associations would have lessened the constraints on chloroplast IEPs, eventually leading to the degradation of ORFs in nearly all group II introns. Although there are at least four protein-assisted pathways for splicing different assemblages of introns in plant chloroplasts, introns belonging to subclass IIA may continue to rely on *matK* (Jenkins et al., 1997; Vogel et al., 1999). The presence of *matK* in the reduced chloroplast genomes of *E. virginiana* and *C. reflexa,* and our example of its preservation after *trnK* intron loss in higher ferns, seem to support this model.

To our knowledge, the higher ferns represent only the third reported case of intron loss with IEP retention — and the first where the intron loss was attributable to an inversion. The expected co-dependence of a group II intron and its IEP makes such an event remarkable. However, plant chloroplast genomes offer an unusual opportunity for IEP isolation due to the seeming dependence of multiple group II introns on a single splice-assisting maturase. The apparent conservation of selection on protein sites suggests that *matK* was performing this generalist function even before this ancient inversion led to the loss of *trnK* and its intron in a fern ancestor. Our study presents a plausible example of a molecule (maturase K) shifting from its primary function as a specific splice factor for the *trnK* intron to a generalist function as a splice factor for multiple group II introns, and finally to being maintained in its new capacity in the plastid genome despite the loss of its co-evolved and previously dependent molecule, the *trnK* intron.

Acknowledgements

Thanks to Dawn Simon, Josh Der, Mark Ellis, and anonymous reviewers for their comments on an earlier draft of the manuscript. This research was funded by National Science Foundation grant DEB-0228432 to PGW, and a Utah State University Vice President of Research fellowship to AMD.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2009.02.006.

References

Barthet, M.M., Hilu, K.W., 2007. Expression of *matK*: functional and evolutionary implications. Am. J. Bot. 94, 1402–1412.

Barthet, M.M., Hilu, K.W., 2008. Evaluating evolutionary constraint on the rapidly evolving gene *matK* using protein composition. J. Mol. Evol. 66, 85–97.

Bonen, L., 1993. *Trans*-splicing of pre-mRNA in plants, animals, and protists. FASEB J. 7, 40–46.

Chapdelaine, Y., Bonen, L., 1991. The wheat mitochondrial gene for subunit I of the *nadH* dehydrogenase complex — a *trans*-splicing model for this gene-in-pieces. Cell 65, 465–472.

Chase, M.W., Cowan, R.S., Hollingsworth, P.M., et al., (19 co-authors), 2007. A proposal for a standardised protocol to barcode all land plants. Taxon 56, 295–299.

Costa, M., Fontaine, J.-M., Loiseaux-de Goër, S., Michel, F., 1997. A group II self-splicing intron from the brown alga *Pylaiella littoralis* is active at unusually low magnesium concentrations and forms populations of molecules with a uniform conformation. J. Mol. Biol. 274, 353–364.

Doyle, J.J., Doyle, J.L., 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. 19, 11–15.

Ems, S.C., Morden, C.W., Dixon, C.K., Wolfe, K.H., dePamphilis, C.W., Palmer, J.D., 1995. Transcription, splicing and editing of plastid RNAs in the nonphotosynthetic plant *Epifagus virginiana*. Plant Mol. Biol. 29, 721–733.

Funk, H.T., Berg, S., Krupinska, K., Maier, U.G., Krause, K., 2007. Complete DNA sequences of the plastid genomes of two parasitic flowering plant species, *Cuscuta reflexa* and *Cuscuta gronovii*. BMC Plant Biol. 7, 45.

Graur, D., Li, W.-H., 2000. Fundamentals of Molecular Evolution, Second ed. Sinauer Associates, Sunderland, Massachusetts.

Hasebe, M., Iwatsuki, K., 1992. Gene localization on the chloroplast DNA of the maiden hair fern; *Adiantum capillus-veneris*. Bot. Mag. Tokyo 105, 413–419.

Hausner, G., Olson, R., Simon, D., Johnson, I., Sanders, E.R., Karol, K.G., McCourt, R.M., Zimmerly, S., 2006. Origin and evolution of the chloroplast *trnK* (*matK*) intron: a model for evolution of group II intron RNA structures. Mol. Biol. Evol. 23, 380–391.

Hess, W.R., Hoch, B., Zeltz, P., Hübschmann, T., Kössel, H., Börner, T., 1994. Inefficient *rpl2* splicing in barley mutants with ribosome-deficient plastids. Plant Cell 6, 1455–1465.

Hilu, K.W., Borsch, T., Müller, K., et al., (16 co-authors), 2003. Angiosperm phylogeny based on *matK* sequence information. Am. J. Bot. 90, 1758–1776.

Hilu, K.W., Liang, H., 1997. The *matK* gene: sequence variation and application in plant systematics. Am. J. Bot. 84, 830–839.

Huang, H.-R., Rowe, C.E., Mohr, S., Jiang, Y., Lambowitz, A.M., Perlman, P.S., 2005. The splicing of yeast mitochondrial group I and group II introns requires a DEAD-box protein with RNA chaperone function. Proc. Natl. Acad. Sci. U. S. A. 102, 163–168.

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755.

Jarrell, K.A., Dietrich, R.C., Perlman, P.S., 1988. Group II intron domain 5 facilitates a *trans*-splicing reaction. Mol. Cell. Biol. 8, 2361–2366.

Jenkins, B.D., Kulhanek, D.J., Barkan, A., 1997. Nuclear mutations that block group II RNA splicing in maize chloroplasts reveal several intron classes with distinct requirements for splicing factors. Plant Cell 9, 283–296.

Kelchner, S.A., 2000. The evolution of non-coding chloroplast DNA and its application in plant systematics. Ann. Mo. Bot. Gard. 87, 482–498.

Kelchner, S.A., 2002. Group II introns as phylogenetic tools: structure, function, and evolutionary constraints. Am. J. Bot. 89, 1651–1669.

Knoop, V., Altwasser, M., Brennicke, A., 1997. A tripartite group II intron in mitochondria of an angiosperm plant. Mol. Gen. Genet. 255, 269–276.

Kosakovsky Pond, S.L., Frost, S.D.W., 2005. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. Mol. Biol. Evol. 22, 478–485.

Kosakovsky Pond, S.L., Frost, S.D.W., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21, 676–679.

Lambowitz, A.M., Perlman, P.S., 1990. Involvement of aminoacyl-transfer RNA-synthetases and other proteins in group-I and group-II intron splicing. Trends Biochem. Sci. 15, 440–444.

Lemieux, C., Otis, C., Turmel, M., 2000. Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. Nature 403, 649–652.

Liere, K., Link, G., 1995. RNA-binding activity of the *matK* protein encoded by the chloroplast *trnK* intron from mustard (*Sinapis alba* L.). Nucleic Acids Res. 23, 917–921.

Magallón, S.A., Sanderson, M.J., 2005. Angiosperm divergence times: the effect of genes, codon positions, and time constraints. Evolution 59, 1653–1670.

Malek, O., Knoop, V., 1998. *Trans*-splicing group II introns in plant mitochondria: the complete set of *cis*-arranged homologs in ferns, fern allies, and a hornwort. RNA 4, 1599–1609.

Matsuura, M., Noah, J.W., Lambowitz, A.M., 2001. Mechanism of maturase-promoted group II intron splicing. EMBO J. 20, 7259–7270.

Michel, F., Umesono, K., Ozeki, H., 1989. Comparative and functional anatomy of group II catalytic introns — a review. Gene 82, 5–30.

Mohr, G., Perlman, P.S., Lambowitz, A.M., 1993. Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. Nucleic Acids Res. 21, 4991–4997.

Moran, J.V., Zimmerly, S., Eskes, R., Kennell, J.C., Lambowitz, A.M., Butow, R.A., Perlman, P.S., 1995. Mobile group II introns of yeast mitochondrial DNA are novel site-specific retroelements. Mol. Cell. Biol. 15, 2828–2838.

Nylander, J.A.A., 2004. MrModeltest v2. Program distributed by the author. Evolutionary Biology Center, Uppsala University.

Posada, D., Crandall, K.A., 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics. 14, 817–818.

Pryer, K.M., Schuettpelz, E., Wolf, P.G., Schneider, H., Smith, A.R., Cranfill, R., 2004. Phylogeny and evolution of ferns (monilophytes) with a focus on the early leptosporangiate divergences. Am. J. Bot. 91, 1582–1598.

Qiu, Y.-L., Palmer, J.D., 2004. Many independent origins of *trans* splicing of a plant mitochondrial group II intron. J. Mol. Evol. 59, 80–89.

Qiu, Y.-L., Libo, L., Wang, B., et al. (21 co-authors), 2006. The deepest divergences in land plants inferred from phylogenomic evidence. 103, 15511–15516.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Roper, J.M., 2007. Structural Evolution of Fern Chloroplast Genomes. Master of Science thesis. Utah State University, Logan, Utah, USA.

Saldanha, R., Chen, B., Wank, H., Matsuura, M., Edwards, J., Lambowitz, A.M., 1999. RNA and protein catalysis in group II intron splicing and mobility reactions using purified components. Biochemistry 38, 9069–9083.

Sanders, E.R., Karol, K.G., McCourt, R.M., 2003. Occurrence of *matK* in a *trnK* group II intron in charophyte green algae and phylogeny of the Characeae. Am. J. Bot. 90, 628–633.

Stein, D.B., Conant, D.S., Ahearn, M.E., Jordan, E.T., Kirch, S.A., Hasebe, M., Iwatsuki, K., Tan, M. K., Thomson, J.A., 1992. Structural rearrangements of the chloroplast genome provide an important phylogenetic link in ferns. Proc. Natl. Acad. Sci. U. S. A. 89, 1856–1860.

Toor, N., Hausner, G., Zimmerly, S., 2001. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. RNA 7, 1142–1152.

Turmel, M., Otis, C., Lemieux, C., 1999. The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the architecture of ancestral chloroplast genomes. Proc. Natl. Acad. Sci. U. S. A. 96, 10248–10253.

Vogel, J., Hübschmann, T., Börner, T., Hess, W.R., 1997. Splicing and intron-internal RNA editing of *trnK-matK* transcripts in barley plastids: support for *matK* as an essential splice factor. J. Mol. Biol. 270, 179–187.

Vogel, J., Börner, T., Hess, W.R., 1999. Comparative analysis of splicing of the complete set of chloroplast group II introns in three higher plant mutants. Nucleic Acids Res. 27, 3866–3874.

Wolf, P.G., Rowe, C.A., Sinclair, R.B., Hasebe, M., 2003. Complete nucleotide sequence of the chloroplast genome from a leptosporangiate fern, *Adiantum capillus-veneris* L. DNA Res. 10, 59–65.

Wolf, P.G., Rowe, C.A., Hasebe, M., 2004. High levels of RNA editing in a vascular plant chloroplast genome: analysis of transcripts from the fern *Adiantum capillus-veneris*. Gene 339, 89–97.

Wolfe, K.H., Morden, C.W., Palmer, J.D., 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc. Natl. Acad. Sci. U. S. A. 89, 10648–10652.

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556.

Yang, Z., 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15, 568–573.

Yang, Z., Bielawski, J.P., 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. 15, 496–503.

Young, N.D., dePamphilis, C.W., 2000. Purifying selection detected in the plastid gene *matK* and flanking ribozyme regions within a group II intron of nonphotosynthetic plants. Mol. Biol. Evol. 17, 1933–1941.

Zimmerly, S., Hausner, G., Wu, X.C., 2001. Phylogenetic relationships among group II intron ORFs. Nucleic Acids Res. 29, 1238–1250.

Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31, 3406–3415.